Phrase Based Statistical Machine Translation Javanese-Indonesian

Aufa Eka Putri Lestari*, Arie Ardiyanti, Ibnu Asror

Scholl of Computing, Infomatics, Telkom University, Bandung, Indonesia

Email: ^{1,*}aufaputry@student.telkomuniversity.ac.id, ²ardiyanti@telkomuniversity.ac.id, ³iasror@telkomuniversity.ac.id

Email Correspondence: aufaputry@student.telkomuniversity.ac.id

Abstract—This research aims to produce a statistical machine translation that can be implemented to perform Javanese-Indonesian translation and to know the influence of the main data sources of statistical machine translation namely parallel corpus and monolingual corpus on the quality of Javanese-Indonesian statistical machine translation. The testing was carried out by gradually adding the quantity of parallel corpus and monolingual corpus to seven configurations of Javanese-Indonesian statistical machine translation. All machine translation configuration experiments were tested with test data totaling 500 lines of Javanese sentences. Results from machine translation are evaluated automatically using Bilingual Evaluation Understudy (BLEU). Test results in seven configurations showed an increase in the evaluation value of the translation machine after the quantity of parallel corpus and monolingual corpus was added. The quantity of parallel corpus in configurations 1 and 2 increased by 3,6%, configurations 2 and 3 increased by 8,23%, configurations 3 and 7 increased by 14,92%. Additional monolingual corpus quantity in configurations 4 and 5 increased BLEU score by 0,18%, configurations 5 and 6 increased by 0,06%, configurations 6 and 7 increased by 0,24%. The test results showed that the quantity of parallel corpus and monolingual corpus could increase the evaluation value of statistical machine translation Javanese-Indonesian, but the quantity of parallel corpus had a greater influence than the quantity of monolingual corpus.

Keywords: Statistical Machine Translation; Parallel Corpus; Monolingual Corpus; BLEU; Phrase Based

1. INTRODUCTION

Humans are individual and social beings that interact with each other. In interacting, language is required to convey purpose and objectives to others. Without language, it can be difficult to convey those purposes and objectives. The function of language for human beings is not only as an intermediary to express themselves, feelings, thoughts, desires, and needs, both as individual and social beings, language is also a tool of integration and social adaptation between humans in developing their civilization [1]. Therefore, language is an important component in everyday life that becomes a means of interacting and communicating between individuals.

Indonesia is one of the countries rich in regional language diversity. The number of regional languages in Indonesia is very large. This is because each province has several regional languages at once. Of the 34 provinces in Indonesia, 718 regional languages have been identified as of 2020. One of the regional languages with the most speakers is Javanese. The Javanese language is a communication media that is primarily spoken by ethnic Javanese people in the central and eastern parts of Java Island. The Javanese language has three levels, namely ngoko, madya, and krama. In everyday life, Javanese ngoko is the Javanese language most commonly used by ethnic Javanese people. The number of regional languages in Indonesia makes not all Indonesians master it. Most Indonesians only speak the regional language used in their region.

The ability to master the language will certainly make it easier to interact and communicate. But some individuals only use certain languages. Machine translation comes as one of the solutions to solve language translation problems. Machine translation is an automatic machine that can process language transfer from one language to another. Machine translation was created to facilitate communication between individuals of different languages.

Machine translation has several approach methods, statistical-based approaches, example-based approaches, and rule-based approaches. Machine translation with statistical approaches is considered capable of overcoming deficiencies from previous approaches, such as rule-based approaches and example-based approaches [2]. Also, in terms of accuracy, machine translation with statistical approaches has better accuracy when compared to machine translation with a rule-based approach [2].

In this research, a phrase-based statistical approach was used in the machine translation Javanese-Indonesian. Statistical machine translation is a type of machine translation where the result is produced based on a statistical model whose parameters are taken from the analysis of parallel corpus [3]. The main idea of the statistical machine translation is that each sentence in the target language is the result of a translation of a sentence from the source language with a certain probability. The translation result is a sentence that has the highest probability. In the construction of the statistical machine translation, the main data source needed is the corpus. Corpus consists of two types, that is parallel corpus and monolingual corpus.

Corpus is a collection of texts both spoken and written in print or electronic media that can be used as a data source. To build a statistical machine translation, parallel corpus containing copies of the source language text and the target language and monolingual corpus containing only text in one language is the target language. Parallel corpus serves to form a translation model, while monolingual corpus serves to form a language model.

The research of regional language statistical machine translation has previously been done by several researchers. Some of these studies include Improving The Accuracy of Statistical Machine Translation by Increasing The Quantity of Monolingual Corpus [4] [5] [6]. The Effect of Parallel Corpus Enhancement on

Statistical Machine Translation Accuracy [7] [8], Indonesian And Javanese Machine Translation Using Phrase-Based Statistical Methods [2] [9], Statistical Machine Translation In Lampung Dialect Api To Indonesian [10], Two-Way Translator Application Web-Based Indonesian-Sambas Malay Language Using Moses Decoder [11], Use of Pivot Language on Statistical Machine Translation English to Sambas Malay Language [12], Algorithm for Sharing Phrases In Sentences To Improve Accuracy of Statistical Machine Translation Indonesian-Bugis Wajo Language [13], and Statistical Machine Translation Accuracy Test (MPS) Indonesian to Sambas Malay And Statistical Machine Translation (Mps) Sambas Malay to Indonesian [1].

The purpose of this research to produce a statistical machine translation that can be implemented to perform Javanese-Indonesian translation and to know the influence of the main data sources of statistical machine translation namely parallel corpus and monolingual corpus on the quality of statistical machine translation Javanese-Indonesian.

2. RESEARCH METHODOLOGY

In general, the development of statistical machine translation is divided into several stages. The initial stage prepares the main data namely parallel corpus and monolingual corpus. Furthermore, parallel corpus and monolingual corpus will be processed through preprocessing, training, decoding, and evaluation stages. Here is the methodology of research statistical machine translation Javanese-Indonesian.

Figure 1. Research Flow

2.1 Creating Corpus

The data used there in this research is a text document of Javanese and Indonesian language taken from online news sites solopos.com, tempo.co, and kompas.com. The Javanese language used is Javanese Ngoko. Furthermore, news text documents in the form of paragraphs are changed into lines of sentences that are inline or called the process of sentence alignment. Sentence alignment is the initial stage of preparing a parallel corpus. Sentence alignment can be seen in Table 1 and Table 2.

Table 1. Parallel Corpus in Paragraph Form

Javanese	Indonesian
Seni Tradisi sing kakandhut ing Kabudayan Jawa iku akeh banget tur werna-werna wujud. Salah sijine yaiku tembang macapat. Lumrahe ditembangke karo rengeng-rengeng diiringi swara gamelan.	Seni Tradisi yang terkandung dalam Kebudayaan Jawa itu sangat banyak dan berbagai macam bentuknya. Salah satunya adalah lagu macapat. Biasanya dilagukan dengan bersenandung diiringan suara gamelan.

Table 2. Sentence Alignment in a Parallel Corpus

Javanese	Indonesian
Seni Tradisi sing kakandhut ing Kabudayan Jawa iku akeh banget tur werna-werna wujudé.	Seni Tradisi yang terkandung dalam Kebudayaan Jawa itu sangat banyak dan berbagai macam bentuknya.
Salah sijine yaiku tembang macapat.	Salah satunya adalah lagu macapat.
Lumrahe ditembangke karo rengeng-rengeng diiringi swara gamelan.	Biasanya dilagukan dengan bersenandung diiringan suara gamelan.

The number of corpus most used in this research was limited to 1350 sentences of a parallel corpus in the Javanese-Indonesian and 5000 sentences monolingual corpus in Indonesian. The limited number of sentences parallel corpus is due to the current lack of Javanese-Indonesian parallel corpus so that the collection is done manually based on Javanese news text sourced from Solopos.com.

2.2 Implementation of Statistical Machine Translation

Building a statistical machine translation Javanese-Indonesian begins with building a system architecture. The architecture of the statistical machine translation Javanese-Indonesian can be seen in Figure 2 below.

Figure 2. Architecture of the Statistical Machine Translation Javanese-Indonesian

Figure 2 shows the design of the statistical machine translation Javanese-Indonesian system which consists of several stages, namely preprocessing, training language model, training translation model, decoding, and evaluation of translation results. At the preprocessing stage parallel corpus and monolingual corpus are processed using Moses software consisting of several stages, tokenization, truecasing, and cleaning.

At the training stage, parallel corpus and monolingual corpus are processed to obtain language model and translation model. Language models are used as a source of text-based knowledge with probabilistic values [14]. Language models produce three types of n-gram models namely unigram, bigram, and trigram. Unigram is the appearance of a word that is not influenced by other words, bigram is the appearance of a word influenced by another word, while trigram is the occurrence of a word influenced by the previous word [9]. Especially for translation model, applied phrase-based approach in which there are several processes, namely alignment, creating phrase table, phrase extracting, phrase score, reordering, and generation models and creating configuration files. At this stage, the process of installing the source language input text with the output text of the target language.

Furthermore, the decoding stage, this stage plays a role to translate the source language to the target language. The decoder will find text in the target language that has the most probability with consideration of the translation model factor and language model [15]. Decoder Moses will translate input sentences in the form of the source language that is Javanese. Furthermore, the input sentence will be processed by the decoder Moses, then from the results of the process will produce the output sentence in the form of sentences from the translation to the target language that is Indonesian.

The last stage is the stage of evaluating the translation results. This evaluation is done automatically. The translation results obtained at the testing stage will be evaluated automatically with BLEU software. The results of the evaluation will be a benchmark of how good the quality of machine translation in this research.

2.3 Testing and Evaluation on Seven Machine Translation Configurations

Testing was carried out to determine the evaluation value of the seven configurations of the machine translation that had been added to the quantity of parallel corpus and its monolingual corpus. Testing and evaluation are done automatically using Bilingual Evaluation Understudy (BLEU). The entire configuration of the machine translation was tested with test data totaling 500 lines of Javanese sentences.

2.4 Analysis of Test Results

Analysis of test results was carried out to determine the influence of the quantity of parallel corpus and monolingual corpus on the quality and evaluation value of statistical machine translation and to identify whether the resulting machine translation was appropriate.

3. RESULTS AND DISCUSSION

In this research, several experiments were conducted on seven configurations of statistical machine translation Javanese-Indonesian by increasing the quantity of parallel corpus and monolingual corpus gradually. The seven configurations of the statistical machine translation are divided into two groups, namely the addition of parallel corpus quantity in configurations 1, 2, 3, and 7 and the addition of quantity to the monolingual corpus in configurations 4, 5, 6, and 7. This is done to determine the influence of parallel corpus and monolingual corpus quantity on the quality and evaluation value of machine translation. The entire configuration of the machine translation was tested with the same test data of 500 Javanese sentences.

3.1 Implementation Statistical Machine Translation Javanese-Indonesian

The first phase of implementation of the statistical machine translation Javanese-Indonesian in this research began with the preprocessing stage. The entire parallel corpus on seven machine translation configurations first through tokenization, truecasing, and cleaning stages. Here is the process of preprocessing the corpus on the statistical machine translation Javanese-Indonesian.

```
aufap@aufap-VirtualBox:~/mt7$ ~/smt/mosesdecoder/scripts/tokenizer/tokenizer.perl -l jv <
~/mt7/train.jv7 > ~/mt7/train.tok.jv7
Tokenizer Version 1.1
Language: jv
Number of threads: 1
```

Figure 3. Tokenization Command

Figure 3 is an order to perform the initial stage of preprocessing namely tokenization. Tokenization is the process of giving the distance between words and also the provision of the distance between words and punctuation [2]. Tokenization is carried out in the corpus of data training and data testing of Javanese and Indonesian. The tokenization process generates a corpus file that has been given the distance between words and between words and punctuation which will then be used in the truecasing process.

```
aufap@aufap-VirtualBox:~/mt7$ ~/smt/mosesdecoder/scripts/recaser/train-truecaser.perl --model ~/mt7/t
ruecase-model.jv7 --corpus ~/mt7/train.tok.jv7
aufap@aufap-VirtualBox:~/mt7$ ~/smt/mosesdecoder/scripts/recaser/train-truecaser.perl --model ~/mt7/t
ruecase-model.id7 --corpus ~/mt7/train.tok.id7
aufap@aufap-VirtualBox:~/mt7$ ~/smt/mosesdecoder/scripts/recaser/train-truecaser.perl --model ~/mt7/t
ruecase-model.mono.id7 --corpus ~/mt7/train.mono.id7
```

Figure 4. Truecasing Command

After the tokenization stage, the tokenization file is then processed to the next stage namely truecasing. Truecasing is a process that plays a role in converting the preceding of each sentence to the most likely place [9]. Figure 4 is a command to perform the truecasing stage. Truecasing is also performed on the training data corpus and data testing corpus.

```
aufap@aufap-VirtualBox:~/mt7$ ~/smt/mosesdecoder/scripts/recaser/truecase.perl --model ~/mt7/truecase
-model.jv7 < ~/mt7/train.tok.jv7 > ~/mt7/train.true.jv
aufap@aufap-VirtualBox:~/mt7$ ~/smt/mosesdecoder/scripts/recaser/truecase.perl --model ~/mt7/truecase
-model.id7 < ~/mt7/train.tok.id7 > ~/mt7/train.true.id
aufap@aufap-VirtualBox:~/mt7$ ~/smt/mosesdecoder/scripts/recaser/truecase.perl --model ~/mt7/truecase
-model.mono.id7 < ~/mt7/train.mono.id7 > ~/mt7/train.true.mono.id7
```

Figure 5. Recasing Command

After the truecasing stage is done, then the recasing stage is carried out using a truecase model that has been made before. Figure 5 is an order to perform the recasing stage.

```
aufap@aufap-VirtualBox:~/mt7$ ~/smt/mosesdecoder/scripts/training/clean-corpus-n.perl ~/mt7/train.true
jv id ~/mt7/train.clean 1 80
clean-corpus.perl: processing /home/aufap/mt7/train.true.jv & .id to /home/aufap/mt7/train.clean, cuto
ff 1-80, ratio 9
Input sentences: 1350 Output sentences: 1350
```

Figure 6. Orders and Results of Corpus Cleaning

The last stage in preprocessing is the cleaning process. Cleaning is a process to eliminate white space and give a limit on sentence length. In this cleaning process, the sentence length is limited to a maximum of 80 words. Limiting the length of sentences is done because the longer the sentence, the more vulnerable the translation errors that will affect the evaluation value. The results of cleaning parallel corpus show that no sentences were wasted, as shown in Figure 6. The number of input sentences is equal to the number of output sentences of 1350 lines. This is because the length of the Javanese parallel corpus sentence is made parallel to the corpus of the Indonesian language partner and there are no sentences whose number of words exceeds 80 sentences.

```
aufa@aufa-VirtualBox:~/mt7/lm7$ ~/irstlm/bin/add-start-end.sh < ~/mt7/train.true.mono.id7> mono.
sb.id7
aufa@aufa-VirtualBox:~/mt7/lm7$ export IRSTLM=$HOME/irstlm;
aufa@aufa-VirtualBox:~/mt7/lm7$ ~/irstlm/bin/build-lm.sh -i mono.sb.id7 -t ./tmp -p -s improved
-kneser-nev -o mono.lm.id7
aufa@aufa-VirtualBox:~/mt7/lm7$ ~/irstlm/bin/compile-lm --text=yes mono.lm.id7.gz mono.lm.arpa
.id7
aufa@aufa-VirtualBox:~/mt7/lm7$ ~/smt/mosesdecoder/bin/build_binary mono.lm.arpa.id7 mono.blm.
id7
```

Figure 7. Language Model Command

After the preprocessing stage is completed, the next stage is the training stage. It is at this stage that the language model and translation model are carried out. The training language model is carried out to get the model of the target language that is Indonesian. Data processed at the training language model stage is a monolingual corpus. Language model are executed using IRSTLM software (The IRST Language Modeling) that has been incorporated into moses. Figure 7 is an order to conduct a training language model. Here's language model generated by IRSTLM.

```
\data\
ngram 1=13071
ngram 2=46182
ngram 3=5570

\1-grams:
-3.83484      pergi      -0.460533
-4.13587      keindahan  -0.204184
-4.67994      belanja   -0.123436
-----
\2-grams:
-1.04666      pemeriksaan kasus -0.061232
-0.566616     saksi kunci   -0.014518
-2.9525       dengan ibu    -0.061232
-----
\3-grams:
-0.210853     aku pergi ke  -0.210853
-1.33483      dengan adanya jaminan
-1.02033      polisi juga meminta
```

Figure 8. IRSTLM Generated Language Model

Figure 8 is a model of the target language that is Indonesian and its probabilistic value. The probability of occurrence of the word sequence 'pergi' in Indonesian is $10^{(-3.83484)} = 1.462$, the occurrence of the word sequence 'dengan' followed by the word 'ibu' in Indonesian has a probability of $10^{(-2.9525)} = 1.115$, and the occurrence of the word sequence 'aku' followed by the word 'pergi' followed by the word 'ke' in Indonesian has a probability of $10^{(0.210853)} = 0.615$. At the language model stage, there are three types of n-gram models, namely unigram, bigram, and trigram. The total number of n-grams used can be seen in Table 3.

Table 3. Number of N-gram Models in Seven Machine Translation Configurations

Machine Translation	Unigram	Bigram	Trigram
Configuration 1	13071	46183	5570
Configuration 2	13071	46184	5569
Configuration 3	13071	46181	5570
Configuration 4	4263	11327	540
Configuration 5	7015	21580	1686
Configuration 6	9396	29750	2037
Configuration 7	13071	46182	5570

It can be seen in Table 3 that at the language model stage obtained the most widely used n-gram model in the entire configuration of the machine translation is bigram. This will cause the language model of the target language that is Indonesian obtained in the decoding process is mostly obtained from the bigram model, where the resulting translation will be influenced by other words.

Furthermore, the training translation model stage, this stage is carried out in the process of installing the source language input text with the target language output text. Translation models in excess using GIZA++ software. Here is the command to do the translation model process.

```

aufap@aufap-VirtualBox:~/mt7/working7$ nohup nice ~/smt/mosesdecoder/scripts/training/train-model.perl
-root-dir train -corpus ~/mt7/train.clean -f jv -e id -alignment grow-diag-final-and -reordering msd-
bidirectional-fe -lm 0:3:$HOME/mt7/lm7/mono.blm.id7:8 -external-bin-dir ~/smt/mosesdecoder/tools >& tr
aining.out &
[1] 2769

```

Figure 9. Translation Model Command

Figure 9 is a command to do the translation model process in building a statistical machine translation Javanese-Indonesian. Translation model generates a moses configuration file, vocabulary corpus, word alignment, and translation model table. Moses configuration files are used to decoding or translating sentences. Once the configuration file is obtained, the translation process can be done.

```

Pengumuman kasebut ||| Pengumuman tersebut ||| 1 0.810811 1 0.697674 ||| 0-0 1-1 ||| 1 1 1 ||| |||
uga nate ||| juga pernah ||| 1 0.275132 0.5 0.611765 ||| 0-0 1-1 ||| 1 2 1 ||| |||
kanggo mbuktekake ||| untuk membuktikan ||| 1 0.838095 1 0.854369 ||| 0-0 1-1 ||| 1 1 1 ||| |||
maneka rupa ||| beraneka ragam ||| 0.5 0.5 0.5 0.0444444 ||| 0-0 1-1 ||| 2 2 1 ||| |||
utawa menehake ||| atau memberikan ||| 1 0.189091 1 0.962963 ||| 0-0 1-1 ||| 1 1 1 ||| |||
bulik lunga menyang ||| tante pergi ke ||| 1 0.039312 1 0.571429 ||| 0-0 1-1 2-2 ||| 1 1 1 ||| |||

```

Figure 10. Pieces of the Javanese-Indonesian Translation Model

Figure 10 is a translation model obtained in the training translation model process. Translation models are pieces of phrases that have been matched between the source language (Javanese) and the target language (Indonesian) that has probability values. The translation model indicates that the probability of translation of the phrase 'maneka rupa' to 'beraneka ragam' is 0.5 or 50%. The language model and translation model will be used on the moses decoder as a machine translation.

After the training stage is completed, the next stage is decoding. Decoding process is carried out to find the text in the target language that has the most probability with consideration of the translation model and language model factor. Here are the commands to do the decoding process.

```

aufa@aufa-VirtualBox:~/mt7/working7$ ~/smt/mosesdecoder/bin/moses -f ~/mt7/working7/mert-work/
moses.ini

```

Figure 11. Decoding Command

Once the command is executed, machine translation can be used immediately. As in Figure 12 decoder moses will translate input sentences in the form of the source language, namely Javanese language. Furthermore, the input sentence will be processed by the decoder moses, then from the results of the process will produce the output sentence in the form of sentences from the translation to the target language that is Indonesian. Here's an example of translation using moses decoder.

```

Aku lunga menyang pasar karo ibu toko daging
Translating: Aku lunga menyang pasar karo ibu toko daging
Line 0: Initialize search took 0.000 seconds total
Line 0: Collecting options took 0.000 seconds at moses/Manager.cpp Line 141
Line 0: Search took 0.047 seconds
Saya pergi ke pasar dengan ibu membeli daging
BEST TRANSLATION: Saya pergi ke pasar dengan ibu membeli daging [11111111] [total=-1
.937] core=(0.000,-8.000,7.000,-2.687,-4.562,-0.890,-1.497,-1.666,0.000,0.000,-1.329,
0.000,0.000,0.000,-59.372)
Line 0: Decision rule took 0.000 seconds total
Line 0: Additional reporting took 0.001 seconds total
Line 0: Translation took 0.049 seconds total

```

Figure 12. Sample Translation Using Moses Decoder

Decoder moses will look for the translation results of the source sentence based on the consideration of the statistical model language model and translation model that has the most probability value. Language models are tasked with finding fluency in translation, while translation models are tasked with finding accuracy in translation.

3.2 Translation Results on Parallel Corpus Quantity Trial

To see the quality of the translation results, a random sample of sentences from the test data was taken in the configuration of the statistical machine translation Java-Indonesia that was tested by increasing the quantity of

parallel corpus gradually. Configuration 1 uses a parallel corpus of 300 sentences, configuration 2 uses 600 sentences, configuration 3 uses 1000 sentences, and configuration 7 uses 1350 sentences. The monolingual corpus used in configurations 1,2,3 and 7 amounts to 5000 sentences.

Table 4. Example of Translation Result of Parallel Corpus Quantity Trial

Input Sentence: Panganan sing legi-legi kerep dadi wedhus ireng penyebab lara untu	
Translation Reference: Makanan yang manis-manis sering menjadi kambing hitam penyebab sakit gigi	
Machine Translation	Translation Results
Configuration 1	Makanan yang legi-legi sering menjadi wedhus ireng penyebab lara untu
Configuration 2	Makanan yang legi-legi sering menjadi wedhus ireng penyebab sakit untu
Configuration 3	Makanan yang legi-legi sering menjadi wedhus hitam penyebab sakit untu
Configuration 7	Makanan yang manis-manis sering menjadi kambing hitam penyebab sakit gigi

In Table 4 it is seen that machine translation configuration 1 is not able to translate the words 'legi-legi', 'wedhus ireng' and 'lara untu'. Machine translation configuration 2 was unable to translate the words 'legi-legi', 'wedhus ireng' and 'untu', but was able to translate the word 'lara' which was not previously able to be translated by configuration 1. Machine translation configuration 3 was unable to translate the words 'legi-legi' and 'wedhus' and 'untu', but was able to translate the word 'ireng' which was not previously able to be translated by machine translation configuration 2. Machine translation configuration 7 is already able to translate all words according to the reference sentence. The inability of machine translation configuration 1, 2, and 3 in translating these words is because the word does not exist in the parallel corpus so that the machine translation detects it as the unknown word.

The success of the machine translation configuration 7 in translating all words according to the translation reference is due to the quantity of parallel corpus used in configuration 7 at most compared to configurations 1,2 and 3. This proves that the more quantity of parallel corpus used, the more translations obtained are also closer to the reference sentence. This is due to the availability of many words and sentence references that can multiply vocabulary in the machine translation.

3.3 Translation Results on Monolingual Corpus Quantity Trial

This test aims to see the influence of monolingual corpus quantity on the quality of translation results, taken random sample sentences from test data on the configuration of the statistical machine translation Javanese-Indonesia tested by increasing the quantity of monolingual corpus gradually. Configuration 4 uses a monolingual corpus of 1000 sentences, 5 configurations use 2000 sentences, configuration 6 uses 3000 sentences and a configuration of 7 uses 5000 sentences. The quantity of parallel corpus used in configurations 4,5,6 and 7 is equal to 1350 sentences.

Table 5. Example of Translation Result of Monolingual Corpus Quantity Trial

Input Sentence: Bathik yaiku lembaran kain sing nduweni motif utawa corak maneka rupa	
Translation Reference: Batik yaitu lembaran kain yang memiliki motif atau corak beraneka ragam	
Machine Translation	Translation Results
Configuration 4	Batik yaitu lembaran kain yang memiliki motif atau corak beraneka ragam
Configuration 5	Batik yaitu lembaran kain yang memiliki motif atau corak berbagai macam
Configuration 6	Batik yaitu lembaran kain yang memiliki motif atau corak berbagai macam
Configuration 7	Batik yaitu lembaran kain yang memiliki motif atau corak berbagai ragam

From table 5 it can be seen that the machine translation configurations 4,5,6 and 7 are already able to translate all words on the input sentence. In Javanese, the word 'maneka' has two meanings, namely 'beraneka' and 'berbagai', while the word 'rupa' has several meanings namely 'ragam', 'macam', and 'wujud'. Translated in configurations 5 and 6, the machine translation found that the match of the words 'maneka' is defined as 'berbagai' and 'rupa' is defined as 'macam', this is because based on the model language the word 'berbagai' has a probability value greater than 'beraneka' and the probability of the word 'macam' is greater than the word 'ragam'. The difference between the translations of the words 'maneka' and 'rupa' in configurations 4,5,6, and 7 are caused by variations in unigram, bigram, and trigram data. Translation errors can also be caused by the absence of unigram, bigram, or trigram data matches in test sentences with unigram, bigram, and trigram data in the parallel corpus, or the word is detected as out of vocabulary.

3.4 Evaluation Value of All Machine Translation Configurations

The evaluation value of the Javanese-Indonesian statistical machine translation can be seen in Table 6 below.

Table 6. Evaluation Value on Seven Configurations

Machine Translation	Parallel Corpus Quantities	Monolingual Corpus Quantities	BLEU Score
Configuration 1	300 sentence	5000 sentence	39,79 %
Configuration 2	600 sentence	5000 sentence	43,39 %
Configuration 3	1000 sentence	5000 sentence	51,62 %
Configuration 4	1350 sentence	1000 sentence	66,06 %
Configuration 5	1350 sentence	2000 sentence	66,24 %
Configuration 6	1350 sentence	3000 sentence	66,30%
Configuration 7	1350 sentence	5000 sentence	66,54 %

Machine translation experiments in configurations 1,2, 3, and 7 were conducted by gradually adding parallel corpus quantities and monolingual corpus was made constant, this was done to find out how much the parallel corpus quantity influenced the evaluation value of the machine translation. The test results listed in Table 6 showed an increase in the BLUE score in configurations 1, 2, 3, and 7. Configurations 1 and 2 improved 3.6%, configurations 2 and 3 improved 8.23%, configurations 3 and 7 increased 14.92%. This proves that the quantity of parallel corpus can improve the evaluation value and quality of the statistical machine translation Javanese-Indonesian. The more quantity of parallel corpus used, the higher the evaluation value.

The next experiment was in configurations 4,5,6, and 7 by increasing the quantity of monolingual corpus gradually, and parallel corpus was made constant, this was done to find out how much influence monolingual corpus quantity had on the evaluation value of the machine translation. The test results in Table 6 showed that the BLEU configuration scores of 4,5,6 and 7 improved. Configurations 4 and 5 increased BLEU scores by 0.18%, configurations 5 and 6 increased BLEU scores by 0.06%, configurations 6 and 7 increased by 0.24%. These results prove that the quantity of monolingual corpus can improve the quality of the statistical machine translation Javanese-Indonesian but the percentage is quite small.

The results of corpus testing on the statistical machine translation Javanese-Indonesian get evaluation value as stated in Table 6. The main factor that affects the quality of statistical machine translation is the quantity of corpus. The higher the number of the corpus used, the higher the evaluation value. The parallel corpus is the biggest factor influencing the evaluation results and quality of statistical machine translation. This is evidenced by the results of experiments in configurations 1, 2, 3, and 7 experienced a greater increase in evaluation value than the increase in evaluation values in configurations 4, 5, 6, and 7.

4. CONCLUSION

From the results of research conducted on the statistical machine translation Javanese-Indonesian obtained the results of statistical machine translation can be implemented to do the translation of Javanese-Indonesian language. Machine translation testing by adding parallel corpus quantity gradually resulted in a machine translation configuration 1 with a BLEU score of 39.79%, configuration 2 obtained a BLEU score of 43.39%, configuration 3 obtained a BLEU score of 51.62% and configuration 7 obtained a BLEU score of 66.54%. From these results, it can be seen that the addition of parallel corpus quantity in configurations 1 and 2 increased by 3.6%, configurations 2 and 3 increased by 8.23%, configurations 3 and 7 increased by 14.92%. The test results prove that increasing the quantity of parallel corpus can improve the quality and evaluation value of the statistical machine translation Javanese-Indonesian. The more quantity of parallel corpus used the better the quality and evaluation value of the statistical machine translation. Machine translation testing by adding monolingual corpus quantity resulted in a machine translation configuration 4 with a BLEU score of 66.06%, configuration 5 obtained a BLEU score of 66.24%, configuration 6 obtained a BLEU score of 66.30% and configuration 7 obtained a BLEU score of 66.54%. From these results there was an increase in BLEU score in configurations 4 and 5 of 0.18%, configurations 5 and 6 increased by 0.06%, configurations 6 and 7 increased by 0.24%. The test results proved that the quantity of monolingual corpus was also proven to increase the evaluation value of the Javanese-Indonesian statistical machine translation. However, because the increase in evaluation value is quite small, it takes a lot of monolingual corpus quantity to get optimal evaluation value.

REFERENCES

- [1] I. Hadi, "Uji Akurasi Mesin Penerjemah Statistik (MPS) Bahasa Indonesia Ke Bahasa Melayu Sambas Dan Mesin Penerjemah Statistik (MPS) Bahasa Melayu Sambas Ke Bahasa Indonesia," *J. Sist. dan Teknol. Inf.*, vol. 2, pp. 1–6, 2014.
- [2] R. Nugroho Aditya, T. Adji Bharata, and B. Hantono S, "Penerjemahan Bahasa Indonesia dan Bahasa Jawa Menggunakan Metode Statistik Berbasis Frasa," *Semin. Nas. Teknol. Inf. dan Komun.*, vol. 2015, no. Sentika, 2015.
- [3] M. Wahyuni, H. Sujaini, and H. Muhandi, "Pengaruh Kuantitas Korpus Monolingual Terhadap Akurasi Mesin Penerjemah Statistik," *J. Sist. dan Teknol. Inf.*, vol. 7, no. 1, p. 20, 2019, doi: 10.26418/justin.v7i1.27241.
- [4] R. Darwis, H. Sujaini, and R. D. Nyoto, "Peningkatan Mesin Penerjemah Statistik dengan Menambah Kuantitas Korpus Monolingual (Studi Kasus : Bahasa Indonesia - Sunda)," *J. Sist. dan Teknol. Inf.*, vol. 7, no. 1, p. 27, 2019, doi:

- 10.26418/justin.v7i1.27254.
- [5] H. Sujaini, "Meningkatkan Peran Model Bahasa dalam Mesin Penerjemah Statistik (Studi Kasus Bahasa Indonesia-Dayak Kanayatn)," *Khazanah Inform. J. Ilmu Komput. dan Inform.*, 2017, doi: 10.23917/khif.v3i2.4398.
 - [6] M. G. Asparilla, H. Sujaini, and R. D. Nyoto, "Perbaikan Kualitas Korpus untuk Meningkatkan Kualitas Mesin Penerjemah Statistik (Studi Kasus : Bahasa Indonesia – Jawa Krama)," vol. 1, no. 2, pp. 66–74, 2018.
 - [7] H. Sujaini, "Peningkatan Akurasi Penerjemah Bahasa Daerah dengan Optimasi Korpus Paralel," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 7, no. 1, 2018, doi: 10.22146/jnteti.v7i1.394.
 - [8] P. Permata, Z. Abidin, and F. Ariyani, "Efek Peningkatan Jumlah Paralel Korpus Pada Penerjemahan Kalimat Bahasa Indonesia ke Bahasa Lampung Dialek Api," *J. Komputasi*, vol. 8, no. 2, pp. 41–49, 2020, doi: 10.23960/komputasi.v8i2.2613.
 - [9] J. Pranata, T. Informatika, F. I. Komputer, and U. D. Nuswantoro, "Mesin penerjemah bahasa indonesia- bahasa jawa 1,2," no. 5, pp. 1–5.
 - [10] P. Permata and Z. Abidin, "Statistical Machine Translation Pada Bahasa Lampung Dialek Api Ke Bahasa Indonesia," *J. Media Inform. Budidarma*, vol. 4, no. 3, p. 519, 2020, doi: 10.30865/mib.v4i3.2116.
 - [11] A. Hidayat, H. Sujaini, and R. Dwinyoto, "Aplikasi Penerjemah Dua Arah Bahasa Indonesia – Bahasa Melayu Sambas Berbasis Web Dengan Menggunakan Decoder Moses," *J. Sist. dan Teknol. Inf.*, 2015.
 - [12] D. W. Ningtyas, H. Sujaini, and N. Safriadi, "Penggunaan Pivot Language pada Mesin Penerjemah Statistik Bahasa Inggris ke Bahasa Melayu Sambas," *J. Edukasi dan Penelit. Inform.*, 2018, doi: 10.26418/jp.v4i2.27414.
 - [13] M. Mulyana, H. Sujaini, and H. S. Pratiwi, "Algortima Pembagian Frasa Dalam Kalimat Untuk Meningkatkan Akurasi Mesin Penerjemah Statistik Bahasa Indonesia – Bahasa Bugis Wajo," *J. Sist. dan Teknol. Inf.*, 2018, doi: 10.26418/justin.v6i2.23984.
 - [14] D. W. Ningtyas, H. Sujaini, and N. Safriadi, "Penggunaan Pivot Language pada Mesin Penerjemah Statistik Bahasa Inggris ke Bahasa Melayu Sambas," *J. Edukasi dan Penelit. Inform.*, vol. 4, no. 2, p. 173, 2018, doi: 10.26418/jp.v4i2.27414.
 - [15] D. Indrayana, H. Sujaini, and N. Safriadi, "Meningkatkan Akurasi Pada Mesin Penerjemah Bahasa Indonesia Ke Bahasa Melayu Pontianak Dengan Part Of Speech," vol. 3, no. 1, pp. 1–5, 2016.